(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 April 2001 (19.04.2001)

PCT

(10) International Publication Number
WO 01/28161 A1

(51) International Patent Classification⁷: H04L 12/413

(21) International Application Number: PCT/US00/27923

(22) International Filing Date: 10 October 2000 (10.10.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/415,901 8 October 1999 (08.10.1999) US

(71) Applicant: MICROSOFT CORPORATION [US/US]; One Microsoft Way, Redmond, WA 98052 (US).

(72) Inventors: VAIDYA, Nitin, H.; 3740 Marielene Circle, College Station, TX 77845 (US). BAHL, Paramvir; 2221 271st Court SE, Issaquah, WA 98029 (US).

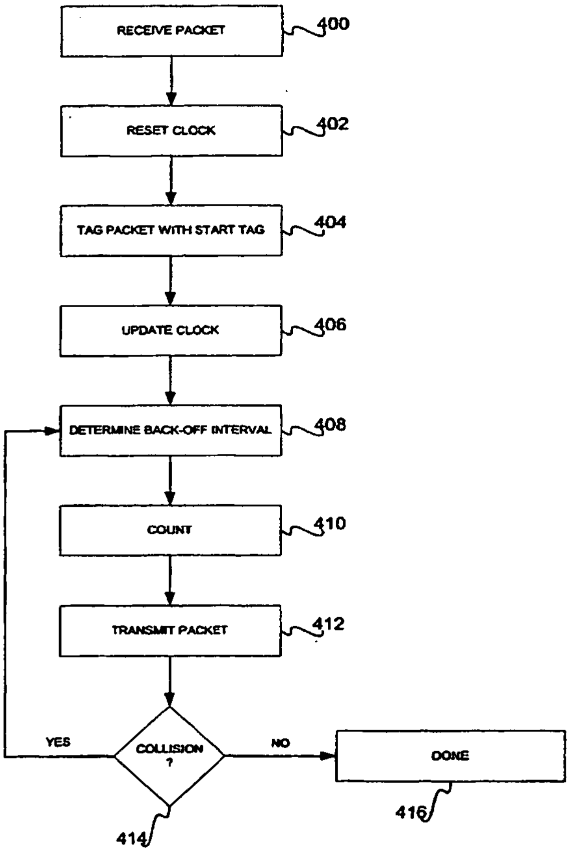
(74) Agent: DRYJA, Michael, A.; Law Offices of Michael Dryja, 704 228th Avenue NE, PMB 694, Sammamish, WA 98074 (US).

(81) Designated States *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: FAIR SCHEDULING IN BROADCAST ENVIRONMENTS



(57) Abstract: Fair scheduling in broadcast environments is disclosed. In one embodiment, a computerized system includes a link through which packets are transmitted, and a plurality of nodes. Each node transmits a packet through the link when counting from a back-off interval reaches a predetermined transmission time. The back-off interval for each packet is based on at least a start tag of the packet, which is assigned to the packet when it arrives at or within the node for transmission over the link, such as at a controller (e.g., a medium-access controller, or MAC) of the node.

WO 01/28161 A1

FAIR SCHEDULING IN BROADCAST ENVIRONMENTS

TECHNICAL FIELD

This invention relates generally to broadcast environments such as wired and wireless networks, multi-hop networks, etc., and more particularly to fair scheduling for data transmission within such environments.

BACKGROUND ART

Broadcast environments include environments in which information is transmitted from discrete originating points over a common medium, and include environments such as networking environments, which have become increasingly common. Networking means that two or more computers or computerized devices, referred to generically as nodes, are communicatively coupled together, so that they can exchange data, typically in the form of packets of data. Networking includes wired local-area-networks (LAN's), in which nodes are connected physically over relatively short distances, wireless LAN's, in which nodes communicate wirelessly over relatively short distances, and multi-hop networks, in which nodes communicate with other nodes on the network, using intermediate nodes to forward their messages..

The amount of data that a network can handle at a given moment in time is referred to as bandwidth. For example, the commonly known Ethernet network generally comes in two different speeds: 100 megabits-per-second (mbps) and 10 megabits-per-second (mbps). This means that, per second, the network is able to accommodate either 100 megabits or 10 megabits of data.

An issue in broadcast environments, such as the ones described above, is determining which node gets to communicate at a given time. Algorithms and schemes to determine which node gets to communicate at a given time typically also concern themselves with fairness. Fairness can be defined in different ways. For example, fairness can mean that each node on the network has a predetermined percentage of the available bandwidth on the network over a given duration of time, a predetermined priority relative to the other nodes on the network, or a weight to divide the available network bandwidth relative to the other nodes. In addition, fairness can mean that a predefined Quality of Service (QOS) is guaranteed for one or more given nodes on the network. A non-restrictive example of QOS is that a given node is guaranteed to receive *x* amount of bandwidth within *y* amount of time after the node requests to transmit data over the network.

FIGs. 2(a)-2(c) are diagrams of example broadcast environments in conjunction with which embodiments of the invention can be practiced;

FIG. 3 is a diagram of an abstraction of a broadcast environment, according to an embodiment of the invention;

FIG. 4 is a flowchart of a method according to an embodiment of the invention; and,

FIG. 5 is a diagram of a computer according to an embodiment of the invention.

MODE(S) FOR CARRYING OUT THE INVENTION

In the following detailed description of exemplary embodiments of the invention, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific exemplary embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be utilized and that logical, mechanical, electrical and other changes may be made without departing from the spirit or scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims.

Some portions of the detailed descriptions which follow are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like. It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout the present invention, discussions utilizing terms such as processing or computing or calculating or determining or displaying or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and elements within the computer 20, such as during start-up, is stored in ROM 24. The computer 20 further includes a hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or writing to a removable optical disk 31 such as a CD ROM or other optical media.

The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive interface 33, and an optical disk drive interface 34, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for the computer 20. It should be appreciated by those skilled in the art that any type of computer-readable media which can store data that is accessible by a computer, such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROMs), and the like, may be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24, or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the personal computer 20 through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port, or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, computers typically include other peripheral output devices (not shown), such as speakers and printers.

The computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as remote computer 49. These logical connections are achieved by a communication device coupled to or a part of the computer 20; the invention is not limited to a particular type of communications device. The remote computer 49 may be another computer, a server, a router, a network PC, a client, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 20, although only a memory storage device 50 has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local-area network (LAN) 51 and a wide-area other node in the network in a direct manner. Example nodes include computers, such as that described in the previous section of the detailed description, as well as computerized devices, such as cell phones, personal digital assistant (PDA) devices, etc.; the invention is not so limited.

Referring next to FIG. 2(c), a diagram of a wireless multi-hop network is shown. The network 230 includes nodes 220, 222, 224, 226 and 228. The nodes communicate with one another in a wireless manner, such as that described in the previous paragraph in conjunction with the description of a wireless LAN. It is noted, however, in the network of FIG. 2(c), that not each node is able to communicate directly with every other node, which is the defining characteristic of a multi-hop network. For example, the node 222 is able to communicate directly with nodes 220 and 224, but not with nodes 226 and 228. Rather, communication between the node 222 and the nodes 226 and 228 must "hop" through node 224. This may be because, for example in the case of wireless communication among the nodes, the node 222 has sufficient communicative range to reach nodes 220 and 224, but not nodes 226 and 228. Example nodes, as before, include computers, such as that described in the previous section of the detailed description, as well as computerized devices, such as cell phones, personal digital assistant (PDA) devices, etc.; the invention is not so limited.

Referring finally to FIG. 3, a diagram of an abstraction of a broadcast environment, such as the broadcast environment of FIG. 2(a), or FIG. 2(b), is shown. In the example, there are flows 304a, 304b, . . . , 304n, which correspond to the nodes of FIG. 2(a), or 2(b). The link 300 corresponds to the network of FIG. 2(a), or 2(b) . The abstraction of FIG. 3 is useful because it shows that when a node, that is, a flow, wishes to transmit a packet of data over a network, that is, a link, the link is commonly shared among all the nodes or flows. Thus, a scheme must be put into place such that all the nodes do not attempt to send packets of data at the same time, else a collision may occur. As described in the background and summary sections, however, the scheme is desirably such that some definition of fairness is achieved as to dividing the bandwidth of the link over the various flows that wish to send data packets thereover.

Method

In this section of the detailed description, a method for distributed fair scheduling of packet transmission among nodes within a network, according to one embodiment of the invention, is described. The method described is distributed in that the method achieves fair scheduling and Quality of Service (QOS) without having a central management node, or other central managing mechanism, coordinating data transmission among the various nodes of the network. As described herein, the method is performed on each node of a network that desires to send packets, such as of data, over the network, as can be abstracted as a link, as described in

In 406, the virtual clock is updated. It is noted that the virtual clock is updated only when a packet is transmitted from a node onto the link. Thus, if at time *t*, a packet is in service, then $v_i(t)$ is updated to

$$v_i(t) = \max\left(v_i(t), s\right)$$

where *s* is the start tag of the packet in service. It is noted that the virtual clock is not updated at any other time in one embodiment.

Once a node i desires to transmit the packet, then in 408, it determines an appropriate back-off interval, which is generally defined as the length of time the node waits until actually transmitting the packet onto the link. This interval is denoted as B_i, and is based on the start tag and the current virtual time $v_i(t)$ in one embodiment. Specifically,

$$B_i = \left\lceil \eta * (S_i^k - v_i(t)) \right\rceil$$

where η, the *Backoff_Multiplier* is a constant in one embodiment.

It is noted that because of the manner in which the start tags and the virtual clock are determined, B_i is non-negative. However, if the start tag and the virtual clock are identical, B_i may become equal to zero. To avoid this, in one embodiment, B_i is further modified as

$$B_i = B_i + X\ ,$$

where X is uniformly distributed in $[1,\ \beta]$ where β, the *Backoff_Window* is a positive integer. This further reduces the probability of back-off intervals of two nodes counting down to zero at the same time. , In 408 as well the *back-off counter* is reset to zero after this step is performed,.

The node then starts counting from the back-off interval to a predetermined transmission time, such as zero, in 410; that is, the node does not actually send the packet until the predetermined transmission time is reached in 410, as counted down from the back-off interval. Thus, in 412, the node has counted down from the back-off interval to the predetermined transmission time, and therefore transmits the packet over the network or link. At this time, the node also tags the packet with a finish tag, as determined as has been described.

In 414, it listens to determine whether another node has sent a packet at exactly the same time, such that a collision resulted. If not, then the method is done in 416. It is noted that in this case, when another packet needs transmission via the method of FIG. 4, that the virtual clock will not be reset in 402, since it is only reset once. However, if a collision has resulted, then a new *back-off interval* (B_i) must be determined, and the packet ultimately resent. Thus, the method goes back to 408. However, in this iteration of 408, the *back-off counter* is increased by function of the form $K_1 * \left(1 - K_2^{-(start\_tag - virtual\_time)}\right)$ is used in one embodiment. In this case, however, after each packet transmission, the back-off counter of each pending packet needs to be redetermined, so as to emulate the scheme in 1).

The method of FIG. 4 can be summarized as follows. Each node of the network may have a pending packet to be transmitted over the network. For each node that does, when the packet actually arrived at the node – for example, when the packet was generated within the node, and then "arrived" at the node for transmission b:/ a medium access controller (MAC) of the node – the packet is stamped with a start tag, and a back-off interval is determined for the node. The nodes thus start counting down from their respective back-off intervals to zero, in one embodiment. The first node that "wins" – the first node that counts down to zero – sends its packet. This node then is able to receive another packet, and the process starts over for that node. Thus, as nodes count down to zero, they fire off their packets. The method also takes care of the situation where more than one node count down to zero at the same time, which results in a collision.

The method of FIG. 4, as one embodiment of the invention, thus provides for fair scheduling within broadcast environments.

Computer

In this section of the detailed description, a computer is described that can function as a node within a network per the method of the preceding section of the detailed description. The computer can be based on, in one embodiment, the computer of FIG. 1, as has been described. However, the invention is not so limited.

Referring now to FIG. 5, the computer includes application programs such as programs 500, 502 and 504. Each of these programs generates data packets, as represented by 506 in the diagram of FIG. 5. The packets are received at, or within, the computer specifically at the controller 508, which can be described in one embodiment as a medium access controller (MAC) – the controller which controls access to the "medium," which in this case is the network, or link. Thus, the method described in the preceding section of the detailed description is performed within the controller 508, to determine when to send a received packet to the link, as represented in FIG. 5 by 510.

Not shown in FIG. 5 is that generally there is a queue to receive packets from the application programs. In such an embodiment, a packet is received by the controller 508 from

We claim:

1. A method for distributed fair scheduling for transmission of packets of data characterized by:
tagging (404) a packet of data with a start tag;
determining (408) a back-off interval based on at least the start tag of the packet;
counting (410) from the back-off interval to a transmission time; and,
transmitting (412) the packet in response to counting from the back-off interval to the transmission time.

2. The method of claim 1, further characterized by:
determining (414) whether a collision occurred between the packet and another packet; and,
in response to determining that a collision occurred, determining (408) a new back-off interval, and transmitting (412) the packet in response to counting from the new back-off interval to a new transmission time.

3. The method of claim 1, further initially characterized by receiving (400) the packet at a node for transmission therefrom.

4. The method of claim 1, further initially characterized by resetting (402) a virtual clock.

5. The method of claim 4, further characterized by updating (406) the virtual clock to the start tag of the packet in response to determining that the start tag exceeds the virtual clock.

6. The method of claim 4, wherein determining (408) a back-off interval is characterized by determining the back-off interval based also on the virtual clock.

7. The method of claim 1, wherein tagging (404) a packet with a start tag is characterized by determining the start tag as set to a greater of a virtual clock and a finish tag of a previous packet.

8. The method of claim 1, wherein the transmission time is zero.

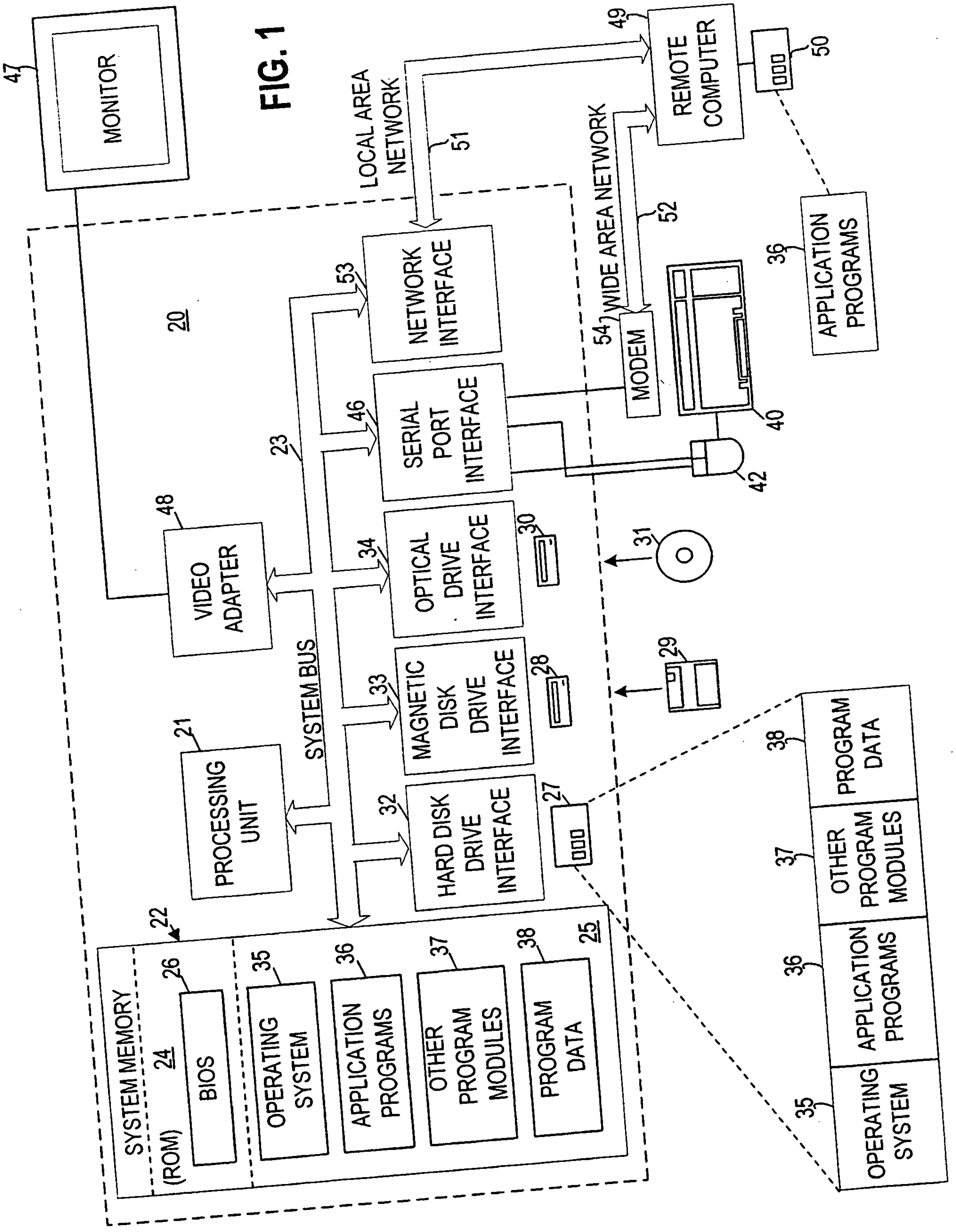

MONITOR
47
20
LOCAL AREA NETWORK
51
49
REMOTE COMPUTER
50
WIDE AREA NETWORK
52
NETWORK INTERFACE
53
MODEM
54
APPLICATION PROGRAMS
36
SERIAL PORT INTERFACE
46
40
42
23
VIDEO ADAPTER
48
OPTICAL DRIVE INTERFACE
34
30
31
SYSTEM BUS
MAGNETIC DISK DRIVE INTERFACE
33
28
29
PROCESSING UNIT
21
HARD DISK DRIVE INTERFACE
32
27
PROGRAM DATA
38
OTHER PROGRAM MODULES
37
APPLICATION PROGRAMS
36
OPERATING SYSTEM
35
22
SYSTEM MEMORY
(ROM)
24
BIOS
26
OPERATING SYSTEM
35
APPLICATION PROGRAMS
36
OTHER PROGRAM MODULES
37
PROGRAM DATA
38
25

FIG. 1

FIG 2(c)

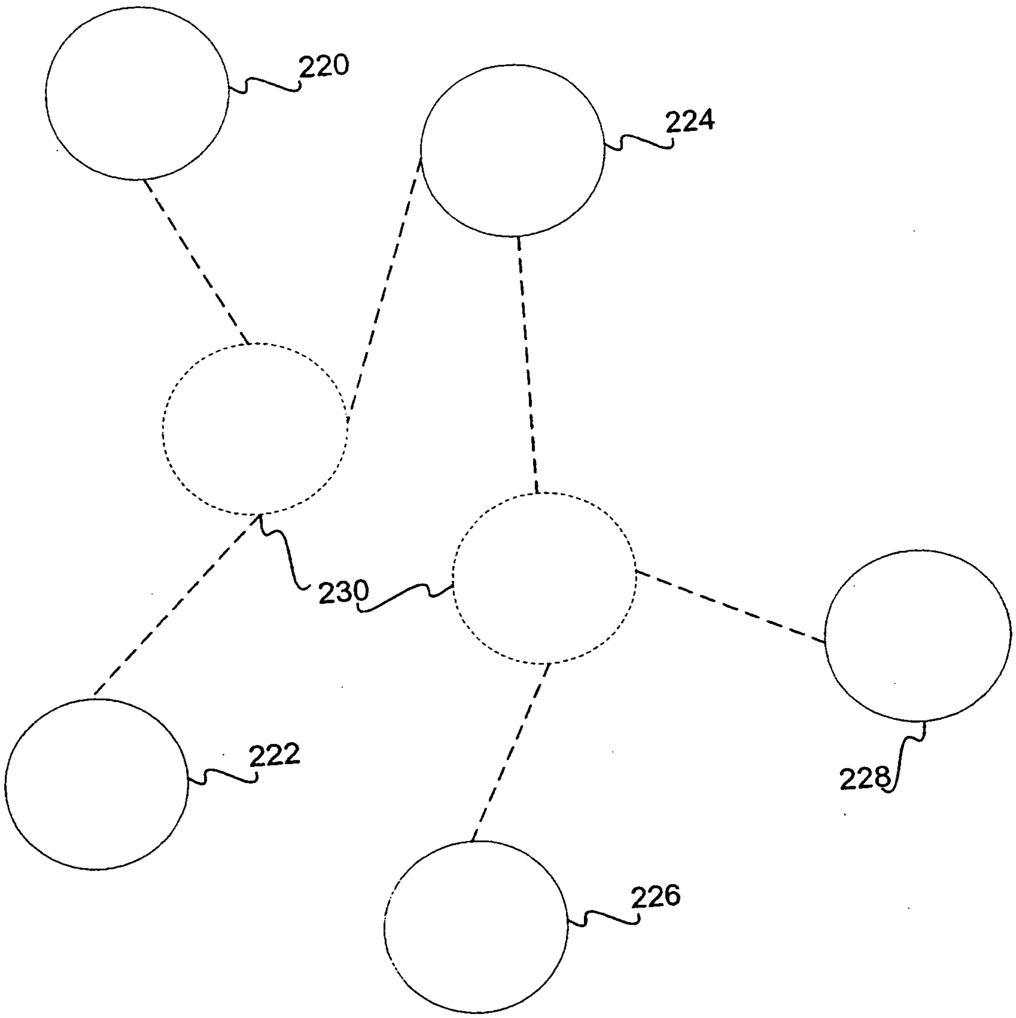
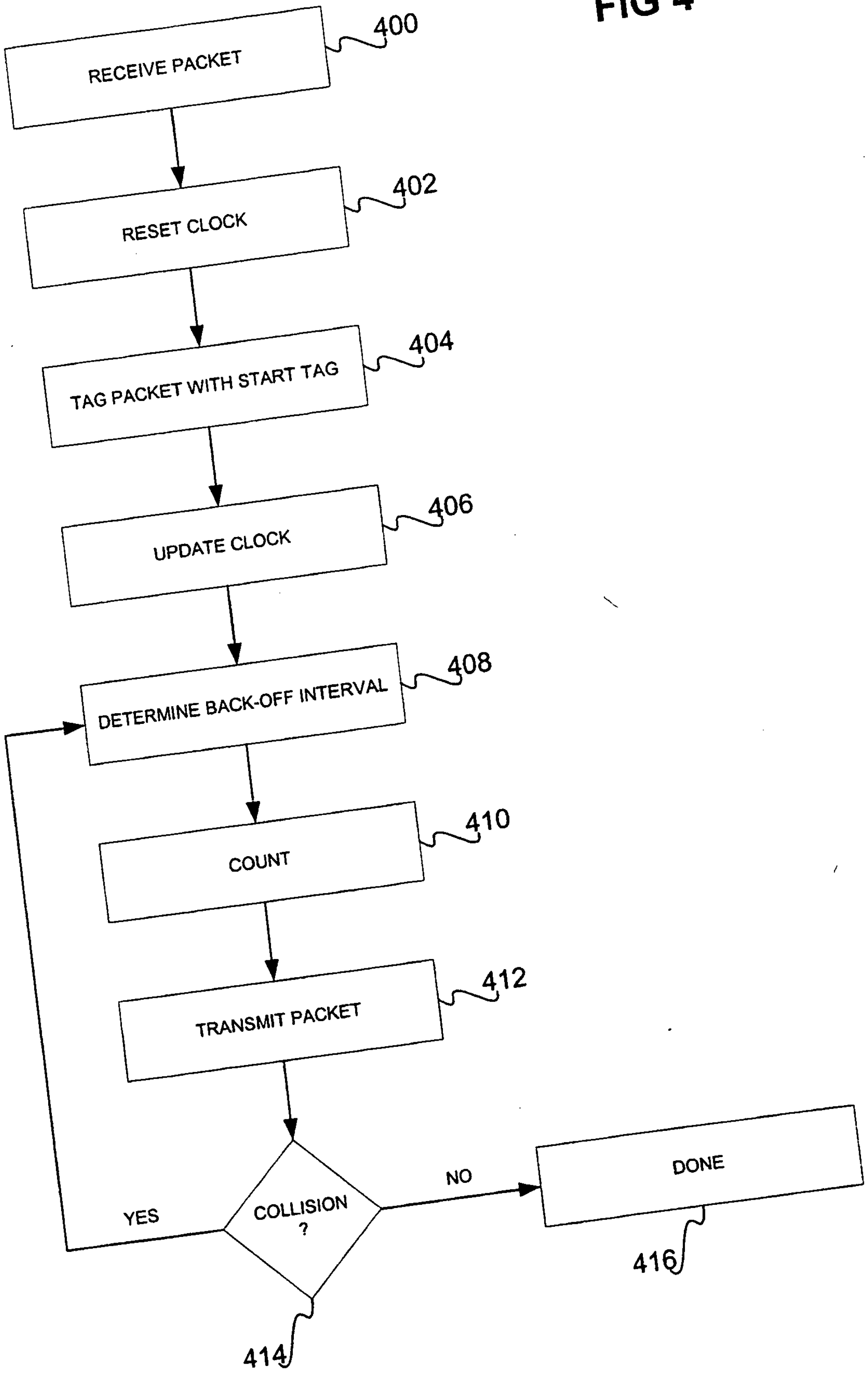


FIG 4


RECEIVE PACKET
400
RESET CLOCK
402
TAG PACKET WITH START TAG
404
UPDATE CLOCK
406
DETERMINE BACK-OFF INTERVAL
408
COUNT
410
TRANSMIT PACKET
412
COLLISION ?
YES
NO
DONE
414
416

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 00/27923

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/413

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, PAJ, INSPEC, IBM-TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 894 559 A (KRISHNA GOPAL ET AL) 13 April 1999 (1999-04-13) abstract column 1, line 33 - line 40 column 3, line 1 - line 23 column 3, line 35 - line 52 column 5, line 52 - line 59 column 6, line 45 - line 61	1-4, 8-16,19, 20
A		5-7,17, 18

-/--

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search 20 February 2001	Date of mailing of the international search report 05/03/2001
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Tous Fajardo, J

INTERNATIONAL SEARCH REPORT

Information on patent family members

ıternational Application No

PCT/US 00/27923

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5894559 A	13-04-1999	EP 0919090 A	02-06-1999
		WO 9807257 A	19-02-1998
		US 6055578 A	25-04-2000
US 5850525 A	15-12-1998	WO 9737464 A	09-10-1997
US 5784375 A	21-07-1998	EP 0904645 A	31-03-1999
		JP 2000512456 T	19-09-2000
		WO 9748209 A	18-12-1997
WO 9301668 A	21-01-1993	DE 4122084 A	07-01-1993
		DE 59208527 D	26-06-1997
		EP 0610202 A	17-08-1994
		JP 6508726 T	29-09-1994
		US 5982781 A	09-11-1999
US 5058108 A	15-10-1991	US 5734659 A	31-03-1998
		US 5621734 A	15-04-1997
		AU 633510 B	04-02-1993
		AU 4141689 A	21-12-1989
		AU 633511 B	04-02-1993
		AU 4141789 A	21-12-1989
		AU 591057 B	30-11-1989
		AU 4266185 A	05-12-1985
		BR 8502706 A	12-02-1986
		CA 1257399 A	11-07-1989
		CA 1279933 A	05-02-1991
		CA 1301941 A	26-05-1992
		DE 3584853 A	23-01-1992
		DE 3586430 A	03-09-1992
		DE 3586430 T	25-03-1993
		DE 3586431 A	03-09-1992
		DE 3586431 T	25-03-1993
		DE 3586433 A	03-09-1992
		DE 3586433 T	08-04-1993
		DE 3586434 A	03-09-1992
		DE 3586434 T	25-03-1993
		DE 3586633 A	15-10-1992
		DE 3586633 T	25-03-1993
		DE 3586634 A	15-10-1992
		DE 3586634 T	01-04-1993
		EP 0163577 A	04-12-1985
		EP 0380141 A	01-08-1990
		EP 0374131 A	20-06-1990
		EP 0374132 A	20-06-1990
		EP 0375664 A	27-06-1990
		EP 0374133 A	20-06-1990
		EP 0374134 A	20-06-1990
		FI 852198 A,B,	02-12-1985
		IE 57544 B	07-10-1992
		JP 2515075 B	10-07-1996
		JP 5063706 A	12-03-1993
		JP 1922242 C	07-04-1995
		JP 6048812 B	22-06-1994
		JP 61056538 A	22-03-1986
		JP 2698336 B	19-01-1998
		JP 8214003 A	20-08-1996
		MX 160504 A	12-03-1990